Variant Calling
Practical Handbook

Summary

During this practical you will

- identify variants
- filter variants

Data Files

This practical will continue from the day3_mapping_BAM_refinement practical. We will use the final bam file created yesterday to perform the variant calling.

bam file: library_final_sorted.bam

reference genome file: Saccharomyces_cerevisiae.EF4.68.dna.toplevel.fa

Are the reference index and dictionary in the same directory as the reference file?

If you are starting your analyses directly from a bam file created by someone else, make sure you have the same reference genome they have used for the alignment. It is essential that the contigs in the reference are the same, in number, length and ID, to those used in the bam file.

Software Used:

Samtools / bcftools are collections of utilities for manipulating sam / vcf files respectively. http://samtools.sourceforge.net/samtools.shtml

Genome Analysis Toolkit (GATK) software is designed for variant discovery and genotyping. http://www.broadinstitute.org/gatk/

Vcftools is a set of scripts to manipulate vcf files. http://vcftools.sourceforge.net/

Getting the Data

Move to your scratch area

```
cd $CINECA_SCRATCH
```

create a directory for today and move into it

```
mkdir day4
cd day4
```

Which is the command to copy here the final bam, and its index and dictionary from yesterday? Where are them?

hint: ../ is the parent directory to the one you are in

VARIANT CALLING

Once the alignments have been refined, snp and indel differences between the data and reference genome can be identified and qualified. Both GATK and samtools are popular softwares to carry out this analysis. Here we will use samtools mpileup.

```
module load profile/advanced
module load autoload samtools
module load autoload bcftools
```

```
samtools mpileup -u -Q 20 -q 50 -g -s -f
../day3/Saccharomyces_cerevisiae.EF4.68.dna.toplevel.fa
library_final_sorted.bam | bcftools call -mv  - > variants_raw.vcf
```

samtools options used:

mpileup generates a bcf or pileup for one or more bam files

-u compute genotype likelihoods and output them in uncompressed binary format (useful for piping commands).

-Q minimum base quality for a base to be considered [default 13]

-q minimum mapping quality for a base to be considered [default 0]

-g generate genotype likelihoods in BCF format

-s output mapping quality

-f faidx-index reference file

Bcftools options used:

call converts between bcf and vcf files

-v output variant sites only

-m multiallelic caller alternative model for multiallelic and rare-variant calling (recommended by samtools)

See samtools manual for more options and details:
http://samtools.sourceforge.net/samtools.shtml

Look at the vcf output file

```
more variants_raw.vcf
```

Try and use the option -t in mpileup. When you check the vcf files, how does this differ from the previous one?

Check the samtools user manual to see other options.

FILTER VARIANTS with VCFTOOLS

The aim of VCFtools is to provide easily accessible methods for working with complex genetic variation data in the form of VCF files. It allows to filter vcf files as well as manipulate them in many useful ways. We are using it here to filter our vcf.

Filters applied:

d=2: minimum coverage 2

w=10: minimum distance from a gap

```
module load vcftools

cat variants_raw.vcf | vcf-annotate -f d=2/w=10 > variants_flt.vcf
```

check your filtered vcf file with the **more** command

If you consider the first ten variants, how many did not pass the filters applied? And why?